Aspect-Aware Image Descriptions for Multimodal Aspect-Based Sentiment Analysis: A Unified Framework with Dual Similarity and Confidence Calibration

Anonymous EMNLP submission

Abstract

Multimodal Aspect-Based Sentiment Analysis (MABSA) involves identifying textual aspects, aligning them with visual evidence, and analyzing their sentiment. Existing approaches often suffer from error propagation and inefficient cross-modal reasoning. To address these challenges, we propose MADSC (Multimodal Aspect-aware Description with Similarity and Calibration) and a unified framework that jointly performs Multimodal Aspect Term EXtraction (MATE), MABSA, and Joint Multimodal Aspect Sentiment Analysis (JMASA) in an end-to-end manner. Firstly, MADSC generates aspect-aware image descriptions by replacing the generic object mentions with textual aspects, bridging the semantic gap between modalities. Second, a dual similarity alignment strategy is proposed to combine textual-object and visual-region alignments using bounding boxes as intermediaries. A confidence calibration mechanism is developed to quantify the uncertainty of alignment, while a modality gating mechanism suppresses irrelevant visual features for absent aspects, ensuring robust predictions. Experiments on benchmark datasets show that MADSC outperforms a wide range of state-of-the-art methods on MATE, MABSA and JMASA tasks.

1 Introduction

Multimodal Machine Learning has become a popular research field due to its ability to incorporate information from multiple modalities (such as text, images, videos, and audio), offering richer representations than Natural Language Processing (NLP). Most representative multimodal tasks involve visual and textual data. These multimodal tasks can be divided into two categories, as illustrated in Fig. 1 and the gap between modalities varies significantly across different types of tasks.

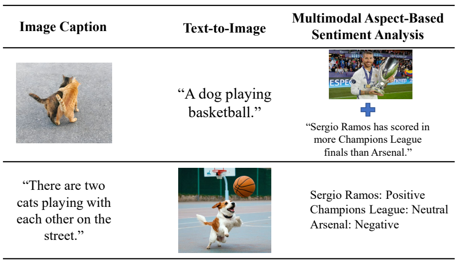


Image Caption
Text-to-Image
Multimodal Aspect-Based Sentiment Analysis
Input
"A dog playing basketball."
"Sergio Ramos has scored in more Champions League finals than Arsenal."
Output
"There are two cats playing with each other on the street."
Sergio Ramos: Positive
Champions League: Neutral
Arsenal: Negative

Figure 1: Example of different tasks.

The first category is referred to as "data-homologous tasks", such as Image Captioning, Text-to-Image and Visual Question Answering, where the information across modalities originates from the same source, so that one modality faithfully translates or extends the other. Thus for data-homologous tasks, the modality gap can be minimized through pre-training to achieve semantic alignment between modalities.

The second category is termed as "data-heterologous tasks" where the information from different modalities is not inherently related. A typical task of this category is Multimodal Aspect-Based Sentiment Analysis (MABSA), which involves identifying textual aspects (typically named entities), aligning them with visual evidence (e.g., visual entities or objects), and analyzing their sentiment polarities. Different from the data-homologous tasks, the discrepancy between modalities in data-heterologous tasks is larger. In MABSA, images and text often exhibit independence. For example, an image may be an extension of the accompanying text or it could be an arbitrary choice by a social media user, so that the visual data may contain a large amount of irrelevant information or may not always contain relevant emotional cues towards an aspect expressed in the text. Consequently, effective modality alignment in data-heterologous tasks, particularly MABSA in this paper, remains a pressing challenge.

Various approaches have been proposed to reduce the inter-modality gap, such as those employed in early models like TomBERT (Yu and Jiang, 2019), which attempt to forcibly map all textual and visual data into a shared feature space. However, this strategy is insufficient for the data-heterologous tasks, as it can inadvertently mix irrelevant information from different modalities, thereby impairing the model's discriminative capability. Traditional methods often fail to distinguish between meaningful visual information and irrelevant visual noise. Instead, they tend to directly concatenate textual and visual features or use shallow attention mechanisms, yet without adequately addressing the potential interference caused by irrelevant modality information (Ju et al., 2021; Yu et al., 2022c; Yang et al., 2022b). To solve the problem in the context of MABSA has involved three closely related subtasks, including Multimodal Aspect Term Extraction (MATE), MABSA, and Joint Multimodal Aspect Sentiment Analysis (JMASA), which will be formulated in Section 3.1. However, existing pipeline approaches often suffer from error propagation and inefficient cross-modal reasoning. They mostly rely on relatively coarse-grained image-text alignment strategies, lacking fine-grained alignment at aspect level.

To address these limitations, we propose a **Multimodal Aspect-aware Description with Similarity and Calibration (MADSC)** method for precise matching between textual aspects and corresponding visual objects, which further facilitates a unified framework that jointly performs MATE, MABSA, and JMASA in an end-to-end manner.

First, to tackle the inherent challenge of aligning discrete textual and continuous visual modalities, MADSC introduces a **dual similarity alignment strategy** that leverages multimodal large language model (MLLM)-generated descriptions for an image as an intermediate modality. MLLMs offer superior cross-modal understanding, generalization, and contextual reasoning capabilities. Such dual similarity alignment strategy helps mitigate the impact of noisy alignments in real-world data, where text-image pairs may not perfectly match (e.g., a text describing food paired with an image of a restaurant exterior).

Furthermore, the alignment is **mediated by visual bounding boxes as intermediaries**, ensuring aspect-related text is accurately linked to visual entities. Unlike prior methods relying solely on direct similarity, our dual strategy introduces an additional alignment pathway via bounding boxes, reducing errors from spurious or missing object mentions in either modality. This approach not only enhances the model's ability to capture aspect-level sentiment but also prevents modality bias, ensuring a robust alignment across modalities and semantic consistency, particularly in tasks requiring fine-grained sentiment analysis.

In addition, cross-modal alignment often suffers from ambiguity, especially when textual aspects lack clear visual counterparts, resulting in unreliable predictions. To address this issue, a **confidence calibration mechanism** is developed to quantify the uncertainty of alignment. By integrating uncertainty estimation, our framework suppresses unreliable multimodal signals, reducing incorrect sentiment classifications caused by misaligned aspects.

Finally, MADSC is incorporated into a framework for joint MATE, MABSA and JMASA. A key challenge is the impact of irrelevant visual information, which introduces noise when the aspects lack visual counterparts. To this end, we propose a **modality gating** mechanism to control the weight of visual information. This ensures that the model relies more heavily on textual features when the aspects are not directly related to the image, to mitigate the unnecessary visual feature interference in data-heterologous tasks, thereby improving the accuracy and robustness of MABSA. Extensive experiments show that MADSC outperforms a range of state-of-the-art methods on the MATE, MABSA and JMASA tasks.

2 Related Work

2.1 Multimodal Aspect-based Sentiment Analysis

Multimodal Aspect-Based Sentiment Analysis (MABSA) has been extensively explored in recent years, focusing on three main subtasks: MATE, MABSA, and JMASA. In MATE, researchers have examined attention-based mechanisms ((Moon et al., 2018; Lu et al., 2018; Zhang et al., 2018)) and Transformer-based architectures ((Yu et al., 2020; Sun et al., 2021; Liu et al., 2022; Zhou et al., 2022; Jia et al., 2023b; Cui et al., 2024)). Additionally, prompt-based learning approaches have been utilized to enhance MATE performance ((Wang et al., 2022; Hu et al., 2023; Li et al., 2023a)). For MABSA, models incorporating cross-modal attention ((Yu and Jiang, 2019; Yu et al., 2019; Zhang et al., 2021b)), multimodal feature fusion techniques ((Yu et al., 2022b; Zhao et al., 2022;

Jia et al., 2023a; Yang et al., 2024)), and auxiliary visual descriptions ((Khan and Fu, 2021; Yang et al., 2022a)) have demonstrated significant improvements. In the realm of JMASA, several integrated models have been proposed to jointly address aspect extraction and sentiment classification, with advancements from methods such as (Ju et al., 2021; Ling et al., 2022; Zhou et al., 2023; Yang et al., 2023a; Peng et al., 2023; Xiao et al., 2024).

2.2 Modality laziness

In multimodal data (especially image-text data), when one modality contains more informative content, the contribution of the other to the outcome is reduced or even ignored. This phenomenon is referred to as Modality Laziness (Du et al., 2023). In some cases, this can be beneficial (for example, when one modality is missing, we can rely on the information from the other modality for inference (Zhao et al., 2021)). However, in most cases, this situation can have a negative impact on the results. Han et al. proposed a method using two bimodal pairs as inputs to address the issue of modality imbalance. Zhang et al. introduced a unimodal optimization approach called MLA, which addresses the issue through alternating unimodal learning. In this paper, we propose a dual similarity calculation method to mitigate this phenomenon, and experimental results show that our method achieves a new state-of-the-art performance.

3 Methodology

3.1 Task Definition and Problem Formulation

We consider three closely related multimodal tasks that integrate both textual and visual information. They are formulated as follows:

Multimodal Aspect Term EXtraction (MATE): MATE aims to identify and classify aspects within the text $T = \{w_1, w_2, ..., w_N\}$ that correspond to visual evidence in the image I. Formally, let $A = \{a_1, a_2, ..., a_M\}$ be the set of aspects, where each aspect a_i spans one or more tokens in T.

Multimodal Aspect-Based Sentiment Analysis (MABSA): MABSA assumes a predefined set of aspect terms and focuses exclusively on classifying the sentiment polarity $s_i \in \{\text{positive}, \text{neutral}, \text{negative}\}$ for each aspect a_i based on the multimodal input (T, I). The final output is a list of sentiment labels associated with the provided aspects, without requiring the model to identify or extract aspect terms.

Joint Multimodal Aspect Sentiment Analysis (JMASA): JMASA aims to extract a set of aspect-sentiment tuples $\{(a_1, s_1), (a_2, s_2), \ldots, (a_m, s_m)\}$, where each a_i is an aspect term span identified from the text and s_i is the corresponding sentiment polarity. JMASA does not assume pre-given aspect terms and requires the model to simultaneously perform aspect extraction and sentiment classification in an end-to-end manner

3.2 Overall Framework of MADSC

Figure 2 sketches the processing flow of **MADSC**. The **Dual-Similarity Module** assigns each aspect–object pair a composite alignment score obtained by combining direct CLIP similarity with an indirect, box-mediated route. These scores are fed into a **Confidence Calibrator**, which converts them into reliability weights. Each weight (i) steers an **Aspect-Aware Caption Generator** that replaces generic object tokens in the MLLM caption with their aligned aspects, and (ii) drives a **Modality Gate** that fuses textual and visual features while attenuating unreliable visual cues. The gated representations are then processed by a **Multimodal Generative Model**, which delivers the predictions for MATE, MABSA, and JMASA tasks. The subsequent subsections elaborate on the design and training objectives of each component.

3.3 Textual Feature Representation and Candidate Aspect Extraction

Given a text sequence $\mathbf{T} = \{w_1, w_2, ..., w_N\}$, we encode each token w_i using a pre-trained language model such as BERT (Devlin et al., 2019):

$$\mathbf{h}_i = \text{BERT}(w_i) \in \mathbb{R}^d \tag{1}$$

where d is the hidden dimension. The encoded sequence is denoted as:

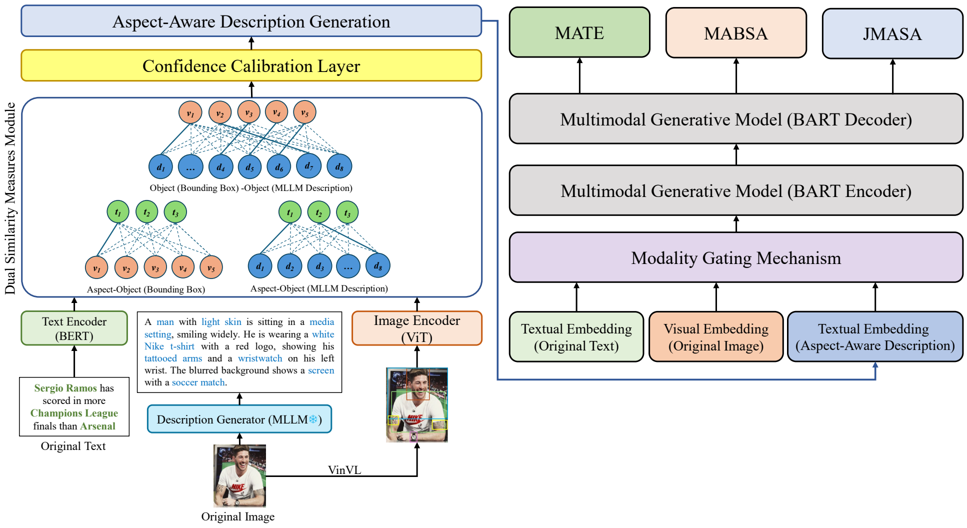
$$\mathbf{H} = [\mathbf{h}_1; \mathbf{h}_2; \ldots; \mathbf{h}_N] \in \mathbb{R}^{N \times d} \tag{2}$$

To identify candidate aspects, we employ the open-source toolkit spaCy[1] to extract aspect spans:

$$A_c = \{a_1, a_2, \ldots, a_M\} \tag{3}$$

where a_i represents the i-th candidate aspect extracted from the text. These candidate aspects are used in the subsequent alignment and sentiment analysis tasks.

[1] https://spacy.io


Aspect-Aware Description Generation
Confidence Calibration Layer
Dual Similarity Measures Module
v_1 v_2 v_3 v_4 v_5
d_1 … d_4 d_5 d_6 d_7 d_8
Object (Bounding Box) -Object (MLLM Description)
t_1 t_2 t_3
v_1 v_2 v_3 v_4 v_5
Aspect-Object (Bounding Box)
t_1 t_2 t_3
d_1 d_2 d_3 … d_8
Aspect-Object (MLLM Description)
Text Encoder (BERT)
Image Encoder (ViT)
Description Generator (MLLM❄)
Original Text
VinVL
Original Image
MATE
MABSA
JMASA
Multimodal Generative Model (BART Decoder)
Multimodal Generative Model (BART Encoder)
Modality Gating Mechanism
Textual Embedding (Original Text)
Visual Embedding (Original Image)
Textual Embedding (Aspect-Aware Description)

A man with light skin is sitting in a media setting, smiling widely. He is wearing a white Nike t-shirt with a red logo, showing his tattooed arms and a wristwatch on his left wrist. The blurred background shows a screen with a soccer match.

Sergio Ramos has scored in more Champions League finals than Arsenal

Figure 2: Framework of our proposed method.

3.4 Visual Feature Representation and Candidate Bounding Box Selection

Given an input image I, we utilize a pre-trained object detection model (VinVL (Zhang et al., 2021a)) to generate region proposals:

$$R = \{r_1, r_2, \ldots, r_L\} \tag{4}$$

where each region r_j includes a bounding box b_j and a detection confidence score c_j. We rank these regions by confidence and select the top K bounding boxes:

$$R_{top} = \{r_1, r_2, \ldots, r_K\} \tag{5}$$

Each selected region r_j is passed through a visual encoder (ViT (Dosovitskiy et al., 2021)) to extract its feature representation:

$$\mathbf{v}_j = \text{ViT}(r_j) \in \mathbb{R}^d \tag{6}$$

The resulting visual feature set is denoted as:

$$\mathbf{V} = [\mathbf{v}_1; \mathbf{v}_2; \ldots; \mathbf{v}_K] \in \mathbb{R}^{K \times d} \tag{7}$$

3.5 Initial Visual Description Generation

Using a MLLM such as GPT4o (OpenAI, 2024), BLIP2 (Li et al., 2023b) or LLaVA (Liu et al., 2024), we can generate a preliminary image description:

$$D_{\text{raw}} = \text{MLLM}(I) \tag{8}$$

where D_{raw} provides a textual summary of the image content, typically including object mentions and basic scene descriptions.

3.6 Dual Similarity Measures and Aspect-aware Description Generation

To align textual candidate aspects $A_c = \{a_1, a_2, \ldots, a_M\}$, visual bounding boxes $R_{top} = \{r_1, r_2, \ldots, r_K\}$, and MLLM-generated visual descriptions D_{raw}, we employ a dual similarity alignment strategy augmented with a confidence calibration mechanism.

Aspect-Bounding Box Similarity. For each candidate aspect $a_i \in A_c$ and bounding box $r_j \in R_{top}$, we compute the multimodal similarity using CLIP (Radford et al., 2021) model:

$$\text{sim}(a_i, r_j) = \cos(\mathbf{t}_{a_i}, \mathbf{v}_{r_j}) \tag{9}$$

where $\mathbf{t}_{a_i} \in \mathbb{R}^d$ is the text embedding of a_i, and $\mathbf{v}_{r_j} \in \mathbb{R}^d$ is the visual feature of r_j.

Object-Bounding Box Similarity. For each object description o_k extracted from D_{raw}, we compute its similarity with each bounding box:

$$\text{sim}(o_k, r_j) = \cos(\mathbf{t}_{o_k}, \mathbf{v}_{r_j}) \tag{10}$$

where $\mathbf{t}_{o_k} \in \mathbb{R}^d$ is the text embedding of o_k.

Aspect-Object Similarity. We directly compute the similarity between each candidate aspect a_i and object description o_k:

$$\text{sim}(a_i, o_k) = \cos(\mathbf{t}_{a_i}, \mathbf{t}_{o_k}) \tag{11}$$

Confidence Calibration for Alignments. To estimate the uncertainty in similarity measurement, we introduce a confidence calibration layer f_{unc}. For each pair (a_i, o_k), it is computed as:

$$u_{a_i, o_k} = f_{\text{unc}}([\text{sim}_{\text{final}}(a_i, o_k)]) \tag{12}$$

where $\text{sim}_{\text{final}}(a_i, o_k)$ is the final similarity score, and f_{unc} is parameterized as:

$$f_{\text{unc}}(\mathbf{x}) = \sigma(W_{\text{unc}} \cdot \mathbf{x} + b_{\text{unc}}) \qquad (13)$$

with $\sigma(\cdot)$ being the sigmoid activation function. This confidence score adjusts the model's reliance on specific alignments by weighting their contributions to downstream tasks.

Final Similarity Score. Each aspect–object pair is compared along two routes: (1)**Direct route**: textual CLIP similarity $\text{sim}(a_i, o_k)$ measures lexical–visual coherence. (2)**Indirect route**: the aspect a_i and object o_k are separately matched to every detected bounding box r_j. The final similarity score between a_i and o_k incorporates both direct and indirect alignments (via bounding boxes). We blend the two routes with learnable coefficients $\alpha, \beta \geq 0, \ \alpha + \beta = 1$:

$$\begin{aligned} \text{sim}_{\text{final}}(a_i, o_k) = \alpha \cdot \text{sim}(a_i, o_k) + \\ \beta \cdot \max_j \left[\text{sim}(a_i, r_j) \cdot \text{sim}(o_k, r_j)\right] \end{aligned} \qquad (14)$$

The adjusted similarity, considering the confidence score u_{a_i,o_k}, is:

$$\text{sim}_{\text{adjusted}}(a_i, o_k) = u_{a_i,o_k} \cdot \text{sim}_{\text{final}}(a_i, o_k) \quad (15)$$

Aspect-Aware Description Generation. Using $\text{sim}_{\text{adjusted}}(a_i, o_k)$, we determine the alignment between aspects a_i and objects o_k in D_{raw}. If $\text{sim}_{\text{adjusted}}(a_i, o_k)$ exceeds a predefined threshold, we replace the object mentioned o_k with its corresponding aspect a_i in the description. For aspects without a valid alignment (i.e., no bounding box or object sufficiently aligned), we append a statement indicating their absence in the image.

3.7 Multimodal Generative Model with Modality Gating

In this step, we integrate all processed information into a multimodal generative model to produce predictions for MATE, MABSA, and JMASA tasks. Additionally, we apply a modality gating mechanism to suppress irrelevant visual contributions for aspects absent from the image.

The model takes three inputs: **Textual features:** The token embeddings $\mathbf{H}$, which encode the original text and candidate aspects. **Visual features:** The bounding box features $\mathbf{V}$, which represent the selected regions. **Aspect-aware description:** The refined description D_{aspect}, encoded as:

$$\mathbf{D}_{\text{aspect}} = \text{BERT}(D_{\text{aspect}}) \in \mathbb{R}^{L \times d}$$

Modality Gating Mechanism. To control the influence of visual features for aspects not aligned with any bounding box, we introduce a gating mechanism. For each aspect–object pair we have already computed a reliability score $u_{a_i,o_k} \in (0, 1)$. We convert this raw confidence into a learnable fusion gate

$$g_{a_i} \ = \ \sigma\big(W_g \, u_{a_i,o_k} + b_g\big) \in (0, 1) \qquad (16)$$

where $W_g, b_g \in \mathbb{R}$ are trainable scalars and $\sigma(\cdot)$ denotes the logistic sigmoid. The final aspect of representation is the convex combination

$$\mathbf{z}_{a_i} \ = \ g_{a_i} \, \mathbf{v}_{a_i} + \big(1 - g_{a_i}\big) \, \mathbf{t}_{a_i} \ \in \ \mathbb{R}^d \qquad (17)$$

Intuitively, high–confidence alignments ($u \to 1$) push the gate towards 1, giving more weight to the visual cue, whereas low–confidence alignments fall back to the textual signal.

Multi-task Outputs. The model generates predictions for three tasks:

1. **MATE:** For each token w_i, predict its aspect label using:

$$P(a_i|T) = \text{softmax}(W_{\text{MATE}} \cdot \mathbf{h}_i + b_{\text{MATE}}) \qquad (18)$$

2. **MABSA:** For each identified aspect a_i, predict its sentiment:

$$P(s_i|a_i) = \text{softmax}(W_{\text{MABSA}} \cdot \mathbf{z}_{a_i} + b_{\text{MABSA}}) \qquad (19)$$

3. **JMASA:** Jointly predict aspect and sentiment using a sequence-to-sequence decoder:

$$P(J|T, I, D_{\text{aspect}}) = \prod_t P(y_t|y_{<t}, \mathbf{H}, \mathbf{V}, \mathbf{D}_{\text{aspect}}) \qquad (20)$$

Loss Function. The model is trained with a multi-task loss:

$$\begin{aligned} L = \lambda_{\text{MATE}} L_{\text{MATE}} + \lambda_{\text{MABSA}} L_{\text{MABSA}} + \\ \lambda_{\text{JMASA}} L_{\text{JMASA}} + \lambda_{\text{conf}} L_{\text{conf}} \end{aligned} \qquad (21)$$

where L_{conf} is the confidence calibration loss:

$$\begin{aligned} L_{\text{conf}} = - \sum_{(a_i,o_k)} [y_{a_i,o_k} \cdot \log u_{a_i,o_k} + \\ (1 - y_{a_i,o_k}) \cdot \log(1 - u_{a_i,o_k})] \end{aligned} \qquad (22)$$

Following the approach in (Ling et al., 2022), we set the trade-off hyperparameters λ_{MATE}, λ_{MABSA}, and λ_{JMASA} to 1, which control the relative contribution of each task and confidence calibration in the multi-task loss function. This multi-task loss encourages accurate aspect recognition, sentiment classification and joint predictions while ensuring the confidence calibration layer effectively modulates uncertain alignments.

4 Experiments

We compare the proposed MADSC with prior methods to answer the following questions: Q1: *Does MADSC effectively bridge the gap between modalities compared to previous methods?* Q2: *Does MADSC achieve state-of-the-art performance on fine-grained image-text recognition tasks?* Q3: *Do the individual modules of MADSC contribute to the improvement of the model's performance?*

4.1 Datasets and Evaluation Metrics

Datasets. We conduct experiments on two multimodal datasets: Twitter-2015 and Twitter-2017 (Yu and Jiang, 2019). In both datasets, each sample contains an image and a piece of text, with one or more aspects. **Evaluation Metrics.** We evaluate the performance of our model on MABSA task by Macro-F1 score (Mac-F1), Accuracy (Acc) while on MATE we use Precision (P), Recall (R) and Micro-F1 score (F1) following previous studies.

4.2 Implementation Details

We set the model learning rate as 5e-5, dropout rate as 0.1, batch size as 16, fine-tuning epochs as 8, and the maximum text length as 256. All the models are implemented on PyTorch with one NVIDIA A6000 GPU. We run our model three times with different random seeds and report the average results. The details of hyperparameter setting are described in Appendix A.

4.3 Baselines

We select a range of competitive baselines for each of the MATE, MABSA, and JMASA tasks.

MATE Baselines. 1) **UMT** (Yu et al., 2020) is the first Transformer-based MATE model. 2) **MAF** (Xu et al., 2022) is a general matching & alignment framework that utilizes the cross-modal matching module to calculate the correlation score between textual and visual modalities. 3) **PromptMNER** (Wang et al., 2022) extracts task-related visual features by a prompt-based visual clue encoder(CLIP). 4) **DGCF** (Mai et al., 2023) is the first MATE model that employed the dynamic cross-modal graph to dynamically construct the interaction of visual and textual nodes. 5) **MNER-QG** (Jia et al., 2023b) leverages queries to acquire prior knowledge about entity categories and visual regions. 6) **PGIM** (Li et al., 2023a) is a two-stage framework employs ChatGPT to generate entity labels by simulating the human cognitive process. 7) **Prompt-Me-Up** (Hu et al., 2023) introduces two novel pre-training tasks to enhance the model's ability to extract entities and relations. 8) **MMIB** (Cui et al., 2024) reduces the visual noises by the modality gating principle and acquires consistent cross-modal representations by an alignment-regularizer.

MABSA Baselines. 1) **TomBERT** (Yu and Jiang, 2019) is the first MABSA model that utillizes BERT to acquire representations. 2) **ESAFN** (Yu et al., 2019) uses attention mechanism to generate aspect-sensitve textual representations. 3) **CapTrBERT-DE** (Khan and Fu, 2021) uses a caption transformer to process images and generate auxiliary sentences. 4) **HIMT** (Yu et al., 2022a) introduces a hierarchical interaction module. 5) SMP (Ye et al., 2022) is a cross-modal contrastive learning module is designed to enhance inter-modality modeling. 6) **VLP-MABSA** (Ling et al., 2022) is the first model that applies the Vision-Language Pre-training model to MABSA. 7) **FITE** (Yang et al., 2022a) utilizes rich facial information to capture visual sentiment cues. 8) **ITOAOF** (Wang et al., 2023) translates images into the input space of the model, alleviating the representation gap between different modalities. 9) **AM-IFN** (Yang et al., 2024) focuses on coarse-grained sentence-image fusion to obtain aspect-guided text-image interaction representations.

JMASA Baselines. 1) **JML** (Ju et al., 2021) introduced a joint MATE and MABSC learning method with an auxiliary cross-modal relationship detection module and a hierarchical framework for visual information processing. 2) **DTCA** (Yu et al., 2022c) proposed a dual-encoder Transformer architecture with tasks for text extraction and visual token matching to improve cross-modal alignment. 3) **CMMT** (Yang et al., 2022b) developed a cross-modal multi-task Transformer with a text-centric cross-modal interaction module to control image influence on text representations. 4) **VLP-MABSA** (Ling et al., 2022) used a unified multimodal encoder-decoder architecture for aspect-sentiment extraction and introduced novel pre-training tasks for Textual and Visual Aspect-Opinion Generation. 5) **AoM** (Zhou et al., 2023) proposed an Aspect-oriented Method with Aspect-Aware Attention and Aspect-Guided Graph Convolutional Network to capture aspect-relevant sentiment. 6) **GMP** (Yang et al., 2023a) used NF-Resnet for image feature extraction and introduced aspect prediction to guide multimodal representation construction. 7) **MOCOLNet** (Mu et al., 2023) proposed a Momentum Contrastive Learning Network that integrates pre-training with the training stage. 8) **MultiPoint** (Yang et al., 2023b) introduced Multimodal Probabilistic Fusion Prompts to improve fusion robustness across different modalities. 9) **DQPSA** (Peng et al., 2023) proposed a framework with Prompt as Query Dual and Energy-based Pairwise Expert modules for aspect-span boundary matching. 10) **Atlantis** (Xiao et al., 2024) integrated image aesthetic assessment for JMASA using a pre-trained model, CoCa for captions, and a High-level RGB-aware Attention Network.

4.4 Main Results

The result on MABSA task is shown in Table 1. On the Twitter2015 dataset, MADSC improves Accuracy by 1.89% and Macro-F1 by 2.67% compared to the state-of-the-art model ITOAOF. On the Twitter2017 dataset, these two metrics are improved by 2.05% and 2.16%, respectively.

Method	Twitter-15		Twitter-17	
	Acc.	Mac-F1	Acc.	Mac-F1
TomBERT, (Yu and Jiang, 2019)	77.15	71.75	70.34	68.03
ESAFN, (Yu et al., 2019)	73.38	67.37	67.83	64.22
CapTrBERT-DE, (Khan and Fu, 2021)	77.92	73.9	72.3	70.2
HIMT, (Yu et al., 2022a)	78.14	73.68	71.14	69.16
SMP, (Ye et al., 2022)	77.53	72.24	71.15	69.47
VLP-MABSA, (Ling et al., 2022)	78.6	73.80	73.80	71.80
FITE, (Yang et al., 2022a)	78.49	73.90	70.90	68.70
ITOAOF, (Wang et al., 2023)	79.45	75.11	74.47	73.05
AMIFN, (Yang et al., 2024)	78.69	75.50	72.29	70.21
MADSC(Ours)	**81.34**	**77.78**	**76.52**	**75.21**

Table 1: Experiment results on MABSA task.

The results on the MATE and JMASA tasks are shown in Table 2 and Table 3 respectively. MADSC demonstrates superior performance across both MATE and JMASA tasks. On the MATE task, MADSC improves F1 score by 2.6% and 3.8% on Twitter-2015 and Twitter-2017 datasets, respectively, compared to state-of-the-art models. On the JMASA task, MADSC outperforms existing methods such as JML and CMMT by 8.8% and 6.4% on Twitter-2015, and 6.0% and 3.5% on Twitter-2017.

Method	Twitter-2015			Twitter-2017		
	Pre.	Rec.	F1	Pre.	Rec.	F1
UMT, (Yu et al., 2020)	71.67	75.23	73.41	85.28	85.34	85.31
MAF, (Xu et al., 2022)	71.86	75.10	73.42	86.13	86.38	86.25
PromptMNER, (Wang et al., 2022)	78.03	79.17	78.60	89.93	90.60	90.26
DGCF, (Mai et al., 2023)	74.76	75.50	75.13	88.50	87.65	88.07
MNER-QG, (Jia et al., 2023b)	77.43	72.15	74.70	88.26	85.65	86.94
PGIM, (Li et al., 2023a)	79.21	79.45	79.33	90.86	92.01	91.43
Prompt-Me-Up, (Hu et al., 2023)	80.03	80.97	80.50	91.97	91.33	91.65
MMIB, (Cui et al., 2024)	74.44	77.68	76.02	87.34	87.86	87.60
MADSC(Ours)	**82.55**	**83.61**	**83.08**	**94.19**	**94.42**	**94.30**

Table 2: Experimental results on MATE task.

Method	Twitter-15			Twitter-17		
	Pre.	Rec.	F1	Pre.	Rec.	F1
JML, (Ju et al., 2021)	65.0	63.2	64.1	66.5	65.5	66.0
DTCA, (Yu et al., 2022c)	67.3	69.5	68.4	69.6	71.2	70.4
CMMT, (Yang et al., 2022b)	64.6	68.7	66.5	67.6	69.4	68.5
VLP-MABSA, (Ling et al., 2022)	65.1	68.3	66.6	66.9	69.2	68.0
AoM, (Zhou et al., 2023)	67.9	69.3	68.6	68.4	71.0	69.7
GMP, (Yang et al., 2023a)	51.6	47.1	49.3	54.2	53.3	53.7
MOCOLNet, (Mu et al., 2023)	66.3	67.9	67.1	67.3	68.7	68.0
MultiPoint, (Yang et al., 2023b)	-	-	66.6	-	-	61.2
DQPSA, (Peng et al., 2023)	71.7	72.0	71.9	71.1	70.2	70.6
Atlantis, (Xiao et al., 2024)	65.6	69.2	67.3	68.6	70.3	69.4
MADSC(Ours)	**72.8**	**73.1**	**72.9**	**72.3**	**71.7**	**72.0**

Table 3: Experimental results on JMASA task.

MATE	**Twitter-2015**			**Twitter-2017**		
	Pre.	Rec.	F1	Pre.	Rec.	F1
MADSC	**82.55**	**83.61**	**83.08**	**94.19**	**94.42**	**94.30**
w/o Confidence Calibration	79.04	78.34	78.69	89.25	91.07	90.15
w/o Modality Gating	77.46	78.20	77.83	87.36	91.18	89.23
replace GPT-4o with BLIP2	80.96	81.14	81.05	91.06	91.68	91.37
replace GPT-4o with LLaVA	81.75	79.66	80.69	91.47	93.64	92.54

Table 4: Results of ablation studies on MATE task.

MABSA	**Twitter-2015**		**Twitter-2017**	
	Acc.	Mac-F1	Acc.	Mac-F1
MADSC	**81.34**	**77.78**	**76.52**	**75.21**
w/o Confidence Calibration	78.62	76.25	74.12	70.98
w/o Modality Gating	77.74	72.64	73.09	70.12
replace GPT-4o with BLIP2	80.52	76.97	75.25	74.09
replace GPT-4o with LLaVA	80.67	77.03	75.66	74.97

Table 5: Results of ablation studies on MABSA task.

4.5 Ablation Study

Tables 4–6 quantify the contribution of each design choice in MADSC. Removing the **confidence calibrator** consistently degrades all three tasks, confirming its role in filtering noisy alignments. Disabling the **modality gate** further reduces performance, indicating that adaptive fusion is preferable to unconditional visual injection. Finally, substituting GPT 4o captions with BLIP2 or LLaVA captions lowers scores across the board, suggesting that caption quality remains a critical factor for robust dual-similarity alignment.

4.6 Case Study

Figure 3 shows a comparison between the predictions from the state-of-the-art model ITOAOF

Image			
Text	This is where [Abe Lincoln]$_{\text{Neu}}$ was not only born , but raised . [Amy Schumer]$_{\text{Neu}}$ at [Lincoln Center]$_{\text{Neu}}$.	[Sergio Ramos]$_{\text{Pos}}$ has scored in more [Champions League]$_{\text{Neu}}$ finals than [Arsenal]$_{\text{Neg}}$.	[Twins]$_{\text{Neu}}$ select [Royce Lewis]$_{\text{Pos}}$ with No.1pick in [MLB]$_{\text{Neu}}$ draft; [Hunter Greene]$_{\text{Neu}}$ to [Reds]$_{\text{Neu}}$.
ITOAOF	Abe Lincoln: Positive(×) Amy Schumer: Neutral(√) Lincoln Center: Neutral(√)	Sergio Ramos: Positive(√) Champions League: Neutral(√) Arsenal: Neutral(×)	Twins: Neutral(√) Royce Lewis: Neutral(×) MLB: Neutral(√) Hunter Greene: Neutral(√) Reds: Neutral(√)
MADSC	Abe Lincoln: Neutral(√) Amy Schumer: Neutral(√) Lincoln Center: Neutral(√)	Sergio Ramos: Positive(√) Champions League: Neutral(√) Arsenal: Negative(√)	Twins: Neutral(√) Royce Lewis: Positive(√) MLB: Neutral(√) Hunter Greene: Neutral(√) Reds: Neutral(√)

Figure 3: Case analysis on ITOAOF and our MADSC model.

JMASA	**Twitter-2015**			**Twitter-2017**		
	Pre.	Rec.	F1	Pre.	Rec.	F1
MADSC	**72.8**	**73.1**	**72.9**	**72.3**	**71.7**	**72.0**
w/o Confidence Calibration	71.3	71.5	71.4	70.9	70.4	70.7
w/o Modality Gating	70.6	70.8	70.7	70.3	70.6	70.5
replace GPT-4o with BLIP2	71.6	71.9	71.8	71.2	71.0	71.1
replace GPT-4o with LLaVA	72.3	71.9	72.1	71.8	71.5	71.7

Table 6: Results of ablation studies on JMASA task.

and our model on three samples. First, in sample (a), our method can correctly predict the neutral sentiment, while ITOAOF makes a wrong prediction. Likewise, in sample (b), there are multiple aspects. Our model correctly excluded the interference from other aspects and predicted a negative sentiment towards "Arsenal." In sample (c), multiple aspects exist and there are strong correlations between them, posing a challenge for previous models. But our model correctly predicts the sentiment for all aspects. These examples demonstrate that our method is effective and can help mitigate the gap between modalities. Appendix B describes the details of the generation of aspect-aware description. Overall, MADSC model demonstrates:

(1) **Enhanced Aspect Contextualisation:** By generating aspect-aware descriptions, the MLLM can provide more nuanced descriptions that incorporate both the visual and textual modalities, offering a comprehensive context for each aspect. This unified aspect representation captures subtle cues from each modality, leading to more accurate and context-aware sentiment predictions.

(2) **Mitigating Modality Bias:** In multimodal sentiment analysis, modality bias—where one modality dominates the sentiment prediction—can reduce model robustness. The MLLM's aspect-aware description generation balances contributions from both modalities, ensuring that the sentiment analysis is grounded in both visual and textual information, and reducing over-reliance on one modality, thereby improving MABSA accuracy.

(3) **Improved Alignment of Aspect-Specific Sentiment:** Aspect-aware descriptions allow for more effective alignment between the visual content and textual descriptions, especially when specific entities are referenced in one modality but not the other. This capability is essential in scenarios where images or text contain modality-exclusive entities, as it minimizes misalignment and supports more accurate aspect-based sentiment recognition.

5 Conclusions

In this paper, we proposed the MADSC model that aims to improve multimodal aspect-based sentiment analysis by effectively aligning textual aspects with visual objects in the image. Through a dual similarity alignment strategy, MADSC generates aspect-aware image descriptions that enhance the accuracy and robustness of three key tasks. It demonstrates superior performance over existing state-of-the-art methods, particularly in handling the fine-grained alignment between text and images and mitigating the impact of irrelevant visual features via the confidence calibration mechanism.

6 Limitations

Despite the encouraging results, several potential avenues for improvement and challenges remain for future research. First, although MADSC effectively handles aspect-aware descriptions in a controlled setting, future work could explore the incorporation of external knowledge bases and prior knowledge during the alignment process to further refine aspect-object relationships. Additionally, while MADSC mitigates the impact of irrelevant visual features, integrating more precise fine-grained attention mechanisms could better capture multimodal dependencies. Moreover, larger and more diverse multilingual datasets should be utilized to evaluate the model's robustness across different domains and real-world scenarios. Finally, leveraging cross-task transfer learning strategies could enhance the model's performance in more complex multimodal settings by utilizing knowledge from multiple subtasks.

References

Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, Tingwen Liu, and Jinqiao Shi. 2024. Enhancing multimodal entity and relation extraction with variational information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL-HLT 2019*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. 2023. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pages 8632–8656. PMLR.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 6–15.

Xuming Hu, Junzhe Chen, Aiwei Liu, Shiao Meng, Lijie Wen, and Philip S Yu. 2023. Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5185–5194.

Li Jia, Tinghua Ma, Huan Rong, and Najla Al-Nabhan. 2023a. Affective region recognition and fusion network for target-level multimodal sentiment classification. *IEEE Transactions on Emerging Topics in Computing*.

Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. 2023b. Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *AAAI 2023*, volume 37, pages 8032–8040.

Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *EMNLP 2021*, pages 4395–4405.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *ACM Multimedia 2021*, pages 3034–3042.

Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023a. Prompting chatgpt in mner: enhanced multimodal named entity recognition with auxiliary refined knowledge. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *ACL 2022*, pages 2149–2159.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Luping Liu, Meiling Wang, Mozhi Zhang, Linbo Qing, and Xiaohai He. 2022. Uamner: uncertainty-aware multimodal named entity recognition in social media posts. *Applied Intelligence*, 52(4):4109–4125.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *ACL 2018*, pages 1990–1999.

Weixing Mai, Zhengxuan Zhang, Kuntao Li, Yun Xue, and Fenghuan Li. 2023. Dynamic graph construction framework for multimodal named entity recognition in social media. *IEEE Transactions on Computational Social Systems*.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *ACL-HLT 2018*, pages 852–860.

Jie Mu, Feiping Nie, Wei Wang, Jian Xu, Jing Zhang, and Han Liu. 2023. Mocolnet: A momentum contrastive learning network for multimodal aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

Tianshuo Peng, Zuchao Li, Ping Wang, Lefei Zhang, and Hai Zhao. 2023. A novel energy based model mechanism for multi-modal aspect-based sentiment analysis. *arXiv preprint arXiv:2312.08084*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *AAAI 2021*, volume 35, pages 13860–13868.

Qianlong Wang, Hongling Xu, Zhiyuan Wen, Bin Liang, Min Yang, Bing Qin, and Ruifeng Xu. 2023. Image-to-text conversion and aspect-oriented filtration for multimodal aspect-based sentiment analysis. *IEEE Transactions on Affective Computing*.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022. Promptmner: prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In *International Conference on Database Systems for Advanced Applications*, pages 297–305. Springer.

Luwei Xiao, Xingjiao Wu, Junjie Xu, Weijie Li, Cheng Jin, and Liang He. 2024. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Information Fusion*, page 102304.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: a general matching and alignment framework for multimodal named entity recognition. In *ACM WSDM 2022*, pages 1215–1223.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022a. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *EMNLP 2022*, pages 3324–3335.

Juan Yang, Mengya Xu, Yali Xiao, and Xu Du. 2024. Amifn: Aspect-guided multi-view interactions and fusion network for multimodal aspect-based sentiment analysis. *Neurocomputing*, 573:127222.

Li Yang, Jin-Cheon Na, and Jianfei Yu. 2022b. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing & Management*, 59(5):103038.

Xiaocui Yang, Shi Feng, Daling Wang, Qi Sun, Wenfang Wu, Yifei Zhang, Pengfei Hong, and Soujanya Poria. 2023a. Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt. In *ACL 2023 Findings*, pages 11575–11589.

Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Soujanya Poria. 2023b. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6045–6053.

Junjie Ye, Jie Zhou, Junfeng Tian, Rui Wang, Jingyi Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Sentiment-aware multimodal pre-training for multimodal sentiment analysis. *Knowledge-Based Systems*, 258:110021.

Jianfei Yu, Kai Chen, and Rui Xia. 2022a. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5408–5414.

Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352.

Yang Yu, Dong Zhang, and Shoushan Li. 2022b. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *ACM Multimedia 2022*, pages 189–198.

Zhewen Yu, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2022c. Dual-encoder transformers with cross-modal alignment for multimodal aspect-based sentiment analysis. In *AACL-IJNLP 2022*, pages 414–423.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021a. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *AAAI 2018*, volume 32.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27456–27466.

Zhe Zhang, Zhu Wang, Xiaona Li, Nannan Liu, Bin Guo, and Zhiwen Yu. 2021b. Modalnet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. *World Wide Web*, 24:1957–1974.

Fei Zhao, Zhen Wu, Siyu Long, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2022. Learning from adjective-noun pairs: A knowledge-enhanced framework for target-oriented multimodal sentiment classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6784–6794.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618.

Baohang Zhou, Ying Zhang, Kehui Song, Wenya Guo, Guoqing Zhao, Hongbin Wang, and Xiaojie Yuan. 2022. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6293–6302.

Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. 2023. AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In *ACL 2023 Findings*, pages 8184–8196.

A Implementation Details

A.1 Hardware and Runtime

Training is conducted on a single NVIDIA RTX A6000 (48 GB) GPU. To keep memory footprint tractable, we *freeze* all vision backbones (**ViT-B/32** and CLIP) and cache their region features, as well as the GPT-4o aspect-aware captions, prior to optimisation. The only trainable components are the BART-based encoder–decoder backbone (150 M parameters) and lightweight task heads, totalling 150.7 M trainable parameters. With batch size 16, one pass over 40 epochs requires ≈5 h 40 m.

A.2 Hyper-parameter Configuration

Table 7 lists all fixed hyper-parameters. Unless noted otherwise, the same setting is used for Twitter-2015 and Twitter-2017.

Hyper-parameter	**Setting**
algorithm	AdamW
learning rate	5×10^{-5}
weight decay	0.01
batch size	16
max length	256 tokens
dropout	0.1
top-K boxes	36
α, β(Twitter2015)	0.7,0.3
α, β(Twitter2017)	0.6,0.4

Table 7: Fixed hyper-parameter settings used in all experiments.

A.3 Pre-processing

- **Region features.** VinVL detects bounding boxes; the highest 36 confidence boxes are encoded by CLIP and cached.
- **Captions.** GPT-4o generates one caption per image. Object tokens aligned to aspects (via Dual-Similarity) are replaced to obtain aspect-aware descriptions.
- **Text normalisation.** All sentences are lowercased and tokenised with the BERT WordPiece tokenizer.

A.4 Sensitivity to the Calibration Weight λ_{conf}

λ_{conf}	MABSA Macro–F$_1$	MATE F$_1$	JMASA F$_1$
0.00	75.90	80.12	70.10
0.25	76.98	81.45	71.25
0.50	**77.78**	82.30	72.10
0.75	77.60	**83.08**	**72.90**
1.00	77.50	83.00	72.80

Table 8: Influence of the calibration loss weight λ_{conf} on the Twitter2015 dataset. Macro–F$_1$ is reported for MABSA; F$_1$ for MATE and JMASA.

Observations. Table 8 and Fig 4 confirms that *confidence calibration is beneficial*: any non–zero λ_{conf} improves the three tasks relative to disabling the term. A moderate weight offers the best trade-off: $\lambda_{\text{conf}} = 0.50$ maximises MABSA, whereas $\lambda_{\text{conf}} = 0.75$ yields the highest joint score for

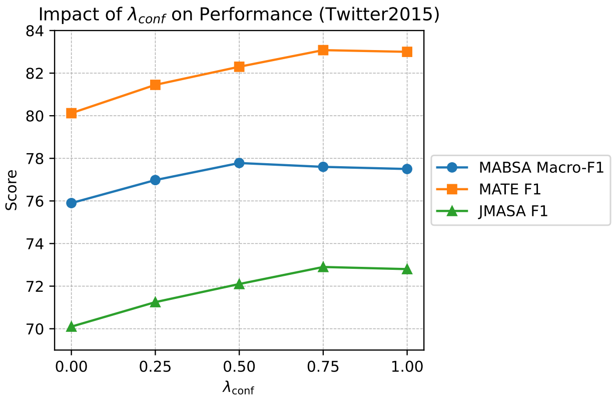
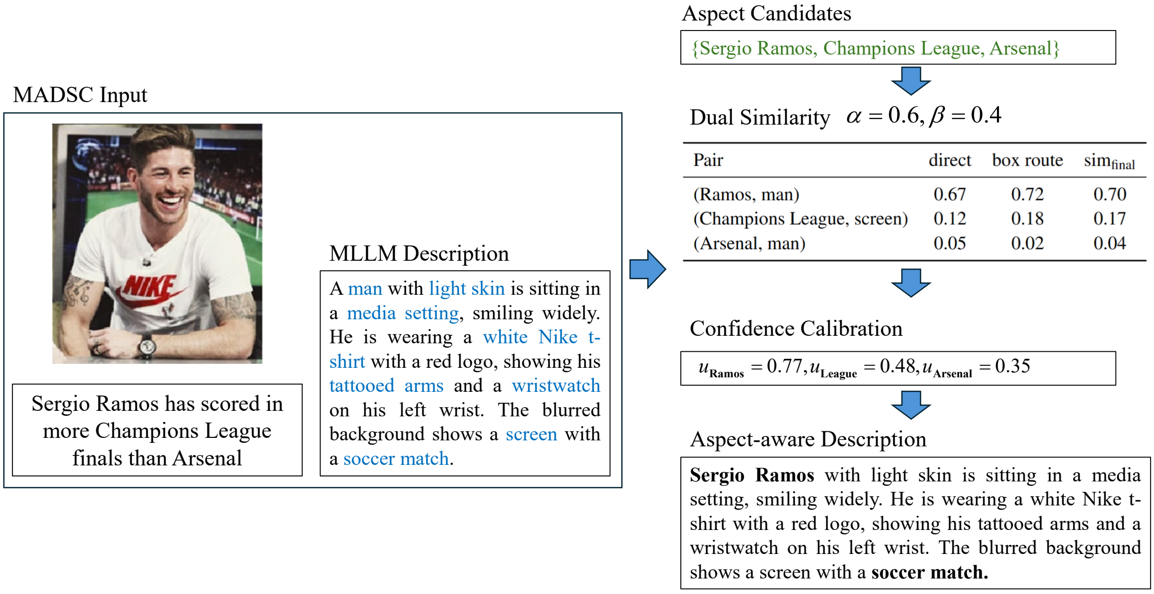


Figure 4: Impact of λ_{conf} on Twitter2015 performance.

MATE and JMASA. Increasing the weight to 1.00 produces no further gains and prolongs training by $\approx 5\%$, indicating diminishing returns.

Recommendation. For datasets with alignment noise comparable to Twitter2015, we advocate selecting $\lambda_{\text{conf}} \in [0.5, 0.75]$. A grid search over $\{0.25, 0.50, 0.75\}$ on the development split usually suffices to find a near-optimal value.

B MADSC Case Analysis


MADSC Input
Sergio Ramos has scored in more Champions League finals than Arsenal
MLLM Description
Aspect Candidates
{Sergio Ramos, Champions League, Arsenal}
Dual Similarity $\alpha = 0.6, \beta = 0.4$
Confidence Calibration
Aspect-aware Description

A man with light skin is sitting in a media setting, smiling widely. He is wearing a white Nike t-shirt with a red logo, showing his tattooed arms and a wristwatch on his left wrist. The blurred background shows a screen with a soccer match.

Pair	direct	box route	$\text{sim}_{\text{final}}$
(Ramos, man)	0.67	0.72	0.70
(Champions League, screen)	0.12	0.18	0.17
(Arsenal, man)	0.05	0.02	0.04

$u_{\text{Ramos}} = 0.77, u_{\text{League}} = 0.48, u_{\text{Arsenal}} = 0.35$

Sergio Ramos with light skin is sitting in a media setting, smiling widely. He is wearing a white Nike t-shirt with a red logo, showing his tattooed arms and a wristwatch on his left wrist. The blurred background shows a screen with a **soccer match.**

Figure 5: Case analysis of aspect-aware description generation in MADSC.

Given the sentence *"Sergio Ramos has scored in more Champions League finals than Arsenal"* and the accompanying image of a smiling male in a white NIKE T–shirt (Fig. 5), the MADSC first identifies three candidate textual aspects—Sergio Ramos, Champions League, and Arsenal. Dual similarity links Ramos to the torso region (*man*) with a high fused score (0.70) and maps Champions League weakly to the TV screen (0.17), whereas Arsenal shows negligible visual correspondence (0.04). The confidence calibrator converts these scores into reliability weights u_{Ramos}=0.77, u_{League}=0.48, and u_{Arsenal}=0.35. During caption rewriting, the object token *man* is replaced by **Sergio Ramos** (high u), while the screen phrase retains its generic form because $u_{\text{League}<0.50}$ falls below the gating threshold. The resulting aspect–aware description therefore reads: *"**Sergio Ramos** is sitting in a media studio, smiling widely. He wears a white Nike T-shirt, displaying tattooed arms and a wristwatch. A TV screen behind him shows a **soccer match**."* This caption explicitly grounds the most reliable aspect in the visual context, provides balanced context for the moderately aligned Champions League, and omits spurious visual cues for Arsenal, thus supplying the downstream sentiment heads with an accurately calibrated multimodal representation.